OXFORD UNIVERSITY PRESS | European Journal of Orthodontics

Which index should be used to measure primary surgical outcome for unilateral cleft lip and palate patients?

Journal:	*European Journal of Orthodontics*
Manuscript ID	EJO-2015-OA-0513
Manuscript Type:	Original Article
Keywords:	Cleft lip / palate, Craniofacial development, Growth

Which index should be used to measure primary surgical outcome for unilateral cleft lip and palate patients?

Structured abstract

Objective: To determine the optimal dentoalveolar measure to assess UCLP patient plaster models.

Design: The models of 34 patients with UCLP taken at 5, 10 and 15-20 years of age were scored by two examiners on two separate occasions using five indices: the 5 Year Olds' (5YO); GOSLON; Modified Huddart/Bodenham (MHB); EUROCRAN and Overjet. Reliability, validity and ease of use were recorded for each index/examiner.

Setting: All models were scored in either Bristol Dental Hospital or Derriford Hospital, Plymouth, United Kingdom by senior orthodontic clinicians.

Results: Highest overall reliability was seen with MHB (Kappa=0.56-0.97). Predictive validity was similar for MHB, GOSLON and 5YO with a 50%-65% prediction of final outcome from 5 and 10 years. EUROCRAN palatal index showed no clear predictive validity (Spearman's correlation=0.20-0.21). Agreement to the gold standard 5YO score at the 5 year age group was high for MHB (Kappa=0.83) and moderate for GOSLON (Kappa=0.59). Agreement to the gold standard GOSLON score at 10 years was highest for 5YO (Kappa=0.69), followed by Overjet (Kappa=0.59) and MHB (Kappa=0.46). Time to score 34 models per index (minutes): GOSLON (13.4) < Overjet (13.6) < 5YO (19.4) < EUROCRAN (24.8) < MHB (27.4).

Conclusion: As an outcome measure of UCLP models, only MHB and 5YO indices can be recommended for use at 5 years of age and GOSLON at 10 years of age.

Introduction

Children born with a cleft of the lip and/or palate face lengthy multidisciplinary treatments for a number of years. The first operation undertaken is typically repair of the lip at around three months of age followed by soft tissue palate repair at six to twelve months of age. These procedures have an immediate impact on both the child and parents by improving feeding, facial aesthetics and ultimately speech development. However, the potential disadvantage is altered maxillary growth (1). Scar tissue formation is an inevitable consequence of surgery and can lead to a restriction of the normal pattern of downward and forward maxillary growth. This can result in complex orthodontic and surgical treatment becoming necessary in the teenage years.

A Europe wide study carried out in the 1980's clearly showed a wide variety in the extent of this maxillary growth restriction following primary cleft surgery (2). The technique and skill of the operator carrying out the surgery is highly likely to have an impact. As with any other field of medicine, audit and a comparison of outcomes between treatment centres allows results to be scrutinised and the overall quality of care to be improved. It is now routine practice to audit the results of primary cleft surgery using one of a number of outcome measures. This is largely done by examining the occlusal records of children with unilateral cleft lip and palate (UCLP) at either five, or ten to twelve years of age. Below is a brief description of the commonly used occlusal outcome measures:

- The GOSLON Yardstick categorises each child's occlusal outcome into one of five categories based on similarity to reference study models (3). This is carried out at ten to twelve years of age.

- The 5 Year Olds' Index (4, 5), also based on a comparison with reference models, but with in depth category descriptions, is applied at five years of age *i.e.* in the primary dentition (see Figure 1).
- The Modified Huddart/Bodenham Index scores each maxillary tooth and its opposing tooth based on the presence and degree of crossbite (6-8). These scores are then summed to produce one overall score. In theory, this allows for finer discrimination between results and also provides a more objective final score. This can be applied to either the five or ten to twelve year age group.
- More recently the EUROCRAN Index has been introduced. This scores palatal morphology as well as the dental arch relationship (9) and a score is assigned for each component from a three point and four point scale respectively. Variants of this index have been developed for application on either the five or nine year age group (9).
- A simple overjet measurement as described by Morris et al. (12).

A more thorough description of all of the above indices can be found in a recent review article (10).

There is little evidence as to which is the most comprehensive outcome measure. As a consequence all are used to some extent, which makes comparison between cleft treatment centres and studies difficult.

This problem has been partly addressed through a recent systematic review of the different indices for assessing primary surgical outcome in UCLP children (11). However, this secondary research is reliant on the available primary data and there is currently no published primary research which attempts to compare these indices.

The aim of the current research was to directly compare the above outcome measures, in order to determine which could be considered to be the most reliable, valid and easy to use.

Method

The indices included in this comparison study were:

- The GOSLON Yardstick
- The 5 Year Olds' Index
- The Modified Huddart/Bodenham Index
- The EUROCRAN Index
- Simple Overjet Measurement

These were chosen based on a general acceptance within the field. Each index has its own set of instructions to follow and these were summarised onto one to two sheets of A4 based on the original references describing their use.

Reference models were already available for the GOSLON Yardstick and 5 Year Olds' Index. Reference models for the EUROCRAN Index were kindly supplied by the developers of this Index. No reference models are required for the Modified Huddart/Bodenham Index or for the measurement of the overjet.

Although some of the indices used in this study can be applied to clefts other than UCLP, the sample only included UCLP in order to simplify the comparison. Similarly the indices were only tested on study models, despite some being validated for use on other media such as photographs as well as study models.

Testing was carried out on both 5 and 10 year study models in order to allow a full comparison of the indices. Final outcome for each patient was also recorded by scoring study models after all treatment had ceased. This treatment may have been orthodontic treatment only or a combination of orthognathic surgery and orthodontic treatment. This ranged from

age 15 to 20 for all included patients. It is possible that further growth could have occurred after final models were taken if taken at 15 years of age, although it would be unlikely for growth after this age to occur to such an extent that it dramatically changed the final occlusion.

In order to collect the study sample, patient records were searched in Frenchay Hospital, Bristol, the Royal United Hospitals, Bath, the Royal Devon and Exeter Hospital, Exeter and Derriford Hospital, Plymouth. To meet the inclusion criteria for the sample, patients had to have presented with a complete UCLP, be non-syndromic, have study models available at 5 and 10 years of age, and post final orthodontic treatment study models at 15 to 20 years of age. Two hundred and eighty three patient records were searched to give a final sample of 34 patients. The following data were collected for each patient: date and types of primary surgery, date the study models were taken, whether or not pre-surgical orthopaedics was carried out, date of orthodontics prior to bone grafting and whether expansion and incisor proclination was carried out, date of secondary alveolar bone grafting and date and type of orthognathic surgery if it was undertaken.

Once these data had been recorded in a Microsoft Excel 2010 spreadsheet, study models for each patient at the three different ages were sent to the laboratory at Frenchay Hospital, Bristol for duplication. These were all duplicated in white stone and identically trimmed to reduce confounding, possible centre identification and therefore bias.

Each set of study models was allocated a random number using a random number generator in Excel downloaded from www.ablebits.com. This number ranged from 1 to 34 for each age group and a suffix was also added to distinguish which age the models were taken at: ‘a’ for 5 years, ‘b’ for 10 years and ‘c’ for final models at 15-20 years. These numbers and letters were added to the models using sticky labels and a database was kept which matched the numbers

to the patient names. The code linking the patient data to the study models was known only to the researcher (TJ).

In order to carry out a fair comparison between the different indices, examiners with similar experience in each index were needed. It proved very difficult to find any examiners with some experience in each index, with most being very experienced in using the GOSLON and 5 Year Olds' indices, but with little experience in using the others. It was therefore decided to use examiners with some cleft experience, but little to no experience with any of the indices. Two consultant orthodontists kindly agreed to be the examiners in Derriford Hospital, Plymouth. A small standardisation exercise was performed on five sets of UCLP models (not included in the main sample) at 5 and 10 years of age. Once complete, a discussion between the examiners and a third party was held to ensure that agreement was reached on the scores for each model using the different indices. This was to reduce systematic bias through any serious misunderstanding of any of the index instructions. It was not designed to make the examiners experts at using each index as this study was partly designed to determine ease of use of each index without prior calibration.

Customised scoring sheets were printed for each age group and index. It was agreed that each of the five indices would be used on 5 and 10 years of age models. The final models taken when the patients were aged 15-20 were only scored using a simplified three point scoring system. This was split into good, moderate and poor and largely based on the opinion of the examiners on the final occlusion, with a poor outcome suggesting a need for orthognathic surgery. These three categories were linked to the five category indices used at younger age groups by a 1 and 2 equating to good, a 3 moderate and 4 and 5 poor. Although this final scale is based on subjective opinion of the examiners, it was designed to be independent of the indices being compared at younger age groups.

Each set of 34 study models at each age group (5, 10 and 15-20 years of age) were arranged in a random order (Figure 2) on large trays so that they could be easily transported. Scoring with each index was carried out on separate scoring sessions to reduce examiner fatigue. The scoring of each model with each index was repeated once by each examiner, leaving a gap of at least one week between first and second scoring sessions. The time taken to score a complete age group with each index was recorded individually using a stop watch. All scoring sessions, including the repeat scoring sessions were completed within one month of beginning the study. After both scoring sessions had been completed for an index, a questionnaire was given to each examiner to provide feedback on the positive and negative aspects of each index.

Once all scoring sessions were completed, the data were entered into an Excel spreadsheet with 10% of the data entry repeated to ensure it had been entered correctly.

Reliability was calculated using weighted Kappa, with a score of 1 indicating perfect agreement and 0 indicating no agreement. Agreement could be calculated for scores recorded between examiners (inter-examiner) and between the same examiner at different time points (intra-examiner) for each index. However, weighted kappa scores can only be calculated for categorical data *i.e.* for GOSLON, 5 Year Olds' and EUROCRAN. A method for converting Huddart/Bodenham to a 5 point scale has been published (8) and was therefore used in this study to allow Kappa to be calculated. A similar conversion has been published for overjet measurement at the ten year age group (12), but none exists at the other age groups.

Validity was measured in two ways. The first by drawing a comparison between each index at each age group to see how closely they categorise the study models to one another. This was done by using one of the indices as the gold standard for each age group. Based on the literature, the gold standard was taken as the 5 Year Olds' index for the five year age group

and the GOSLON Yardstick for the ten year age group. Although there is no high level evidence confirming these indices are the best at these age groups, this was based on the frequency these indices were used in the literature at their respective age groups as well as expert opinion. This enabled the number and complexity of the statistical tests used to be greatly reduced meaning the likelihood of finding a true difference between the indices' validity was increased. The other indices which could be converted to a matching five category scale were compared directly to the gold standard by calculating weighted kappa scores *i.e.* how well they agreed with the gold standard scores. EUROCRAN has a different number of groups and so could not be compared in this way.

An alternative method of measuring validity is to look at the predictive validity for each index. A comparison was drawn between results of the different indices at the 5 and 10 year age groups, and the outcomes of the final study models taken at around 20 years of age. The 20 year age group scores recorded using the GOSLON Yardstick were simplified into poor, moderate and good outcomes. As previously described, those indices used at the five and ten year age groups with similar categories to GOSLON could also be converted to poor, moderate and good outcomes for comparison (any patient who required orthognathic surgery was recorded as a poor outcome for the purposes of this study). This comparison was carried out by calculating the percentage of study models which stayed in the same category scored in the younger age group, compared to the final age group, together with the percentage which improved and the percentage which worsened.

A formal institutional review board approval is not available for this study as it was not considered necessary. The principles outlined in the Declaration of Helsinki were followed throughout this study.

Results

A final sample of 34 UCLP patient's records were included for the index comparison study. Of these, 11 (32.4%) were female and 23 (67.6%) were male and Table 1 summarises the age range within each group.

Time taken when scoring with each index

The mean time with 95% reference ranges taken to score each set of 34 study models with each individual index for both the 5 and 10 year old records can be seen in Figure 3.

This shows that both the GOSLON Yardstick and overjet measurement took the shortest time to use and the EUROCRAN and Modified Huddart/Bodenham took the longest.

Reliability results

Intra- and inter-examiner reliability results can be seen in Table 2 and 3 respectively.

Substantial (0.6-0.8) or almost perfect (>0.8) intra-examiner agreement was achieved for all indices apart from one GOSLON five year old and one EUROCRAN dental component, which both achieved moderate agreement (0.4-0.6).

The modified Huddart/Bodenham index had the best inter-examiner reliability with almost perfect agreement. The 5 Year Olds' Index, overjet measurement and EUROCRAN dental component managed substantial agreement. GOSLON also achieved substantial agreement other than one score at 5 years of age, which had moderate agreement. EUROCRAN palatal only managed moderate inter-examiner reliability at the 5 year age group.

Validity results based on comparison to gold standard

Face validity was measured by comparison of each index to a gold standard index, namely the 5 Year Olds' Index at the 5 year age group and the GOSLON Yardstick at the 10 year age group.

Figures 4 and 5 show that the modified Huddart/Bodenham index achieved almost perfect agreement with the gold standard at the 5 year age group, but dropped to moderate agreement at the 10 year age group. GOSLON agreement at the 5 year age group to the gold standard was moderate and the 5 Year Olds' Index at the 10 year age group was substantial. Although the EUROCRAN index cannot be compared in the same way, Spearman's correlation coefficients could be calculated to show whether there is a correlation between the gold standard scores. This is less optimal than calculating agreement as it does not necessarily confirm a high agreement between individual scores for the same or different examiners, but it gives an idea of correlation between the indices. EUROCRAN dental achieved a Spearman's coefficient of 0.9 (p value<0.001). EUROCRAN palatal showed little correlation with values of 0.27 (p value$=0.13$) at the 5 year age group and -0.05 (p value$=0.77$) at the 10 year age group. Overjet measurement at the 5 year age group showed a strong inverse correlation with a Spearman's coefficient of -0.91 (p value<0.001).

Predictive validity results

Association between scores recorded at five and ten year age groups and scores at the final 15-20 age group can be tested by looking at the percentage of these scores which stay the same across the age groups. The percentage which get better and the percentage which get worse. A high percentage, which stays the same between the initial score and the score

recorded for the final age group indicates a high predictive validity for the index. The final outcome of the 20 year age group was graded as good, moderate or poor and the scores for the 5 and 10 year age group were also converted to this grouping where possible.

Table 4 shows that correct prediction of final grouping was around 50% for GOSLON, 5 Year Olds' and modified Huddart/Bodenham at the 5 year old age group. This increased to 60-65% for the 10 year age group. Overjet measurement was lower at the 10 year age group with 44% correct prediction of final grouping. Spearman's correlation coefficients can be calculated for EUROCRAN dental and palatal components (and overjet measurement at the 5 year age group) in a similar way as described for the previous validity section. These values can be seen in Table 5.

Ease of use questionnaire results

The ease of use questionnaires given to examiners once they had finished the scoring sessions can be seen in Appendix 1. Table 6 shows the average subjective scores for each index assigned by the examiners for ease of use. This is based on a score from 1-10 with 1 being very easy to use and 10 being very difficult to use.

Discussion

Ease of use of indices

In order for an outcome measure to become widely adopted in a busy clinical environment, it is essential that it is easy to use. This includes time taken to complete scoring and the tools necessary for scoring e.g. reference models, training required and user friendliness.

The times taken for scoring models were recorded per age group for each index. Although the mean time taken shows a clear difference between the indices, there is significant overlap as illustrated by the 95% reference ranges, meaning the difference in time taken to complete scoring is less clear cut (Figure 4). There was certainly a trend for scoring with both the Modified Huddart/Bodenham and EUROCRAN indices to take the longest to complete. This is not surprising since individual teeth are scored together with their opposing tooth and all scores need to be added up when using the Modified Huddart/Bodenham index. This is time consuming and does not include the time taken to convert the final score into a five category scale similar to the GOSLON or 5 Year Olds' Index. The EUROCRAN Index not only requires the dental component to be scored, but also the palatal morphology, which would account for the increased time taken to complete scoring.

It is perhaps more surprising that the 5 Year Olds' Index appears to take longer to score models than the GOSLON Yardstick. This may be because the description for each category used in the GOSLON Yardstick is not as thorough or lengthy as for the 5 Year Olds' Index. Although this is likely to make the GOSLON Yardstick more subjective and possibly negatively impact reliability and validity, it can act as a positive influence on time taken to use the index. With less complex descriptions for each group, a snap decision needs to be made based on the evidence available leading to a shorter overall scoring time. However, it is possible that with experience the time taken for the 5 Year Olds' Index would shorten, with easily categorised models being scored quickly and the in depth category descriptions only being re-read for more difficult sets of study models. The GOSLON Yardstick along with the simple overjet measurement were the quickest of all the indices to use.

Other aspects of user friendliness are very similar for both the 5 Year Olds' Index and the GOSLON Yardstick. Both require a set of reference models and ideally a calibration course. Feedback from the examiners confirms that they would find calibration useful and felt more

experience in their use would probably lead to improved reliability. Both examiners felt that the 5 Year Olds' Index was more objective compared with GOSLON, but with the trade-off that it required careful examination of the category descriptions and was therefore more time consuming to use. The examiners commented that both indices accurately reflected their own thoughts on 'how the case looked', although much more 'guessing' was needed when using the 5 Year Olds' Index at 10 years of age and the GOSLON Yardstick at 5 years of age.

The EUROCRAN Index required the most reference models and the lengthiest description for examiners to read. Although the examiners felt the index was thorough, it was the most complex and was rated as the most difficult of all the indices to use. The dental base relationship was commented on as being difficult to assess by one examiner, while both found the palatal morphology challenging to categorise (confirmed by the reliability scores).

Overjet measurement was rated as the easiest outcome measure of all to use. This required very little training or instruction and no reference models. However, both examiners questioned its validity with it having no input from either transverse or vertical discrepancies. Some practical problems were also noted such as worn incisal edges, anterior open bites and the absence of incisors, all making accurate overjet measurement more difficult.

Finally, the Modified Huddart/Bodenham index was also highly rated for ease of use and had very positive overall feedback. Although it was time consuming and did require some arithmetic (which could lead to mistakes in the final score), it was described as the most objective index with both examiners rating it as having the least margin for error in assigning scores. No reference models are required and very little training seems to be needed, although it should be noted that both examiners were very experienced orthodontic consultants and therefore used to assessing occlusions in detail. Its use without training by someone less experienced or in a different field may not produce such favourable results. One examiner

questioned its validity due to the lack of vertical discrepancies informing the final score, while the other felt that the incisor weightings were too low, meaning that A-P discrepancies could sometimes be under-rated.

Reliability

Testing the reliability of an index is relatively straight forward if scoring is repeated with multiple examiners. Reliability is clearly likely to be affected by an examiner's familiarity with a particular outcome measure. It was decided that neither examiner should go on any calibration course prior to taking part in this study, contrary to the advice for some of the indices used. As calibration courses only exist for some of the indices being tested, it was felt that this may make the examiners unfairly familiar with those indices which could introduce bias into the results.

The inter-examiner and intra-examiner reliability scores as assessed using weighted Kappa showed some interesting trends. It is important to first note what is considered as acceptable reliability in the wider literature. Published weighted Kappa scores have a wide range from 0.56 to 1.00 (7, 13-17). Ideally, in large multi-centre studies intra-examiner weighted Kappa scores should be >0.8 and inter-examiner >0.7 to ensure results are reliable.

In the present study the GOSLON Yardstick weighted Kappa scores were around 0.8 for intra-examiner (although with a large range) and broadly less than 0.7 for inter-examiner reliability. This is perhaps slightly worse than expected. There are several potential explanations for this. Both examiners began the study by using this index, so it may have performed slightly worse whilst they became familiar with the method of scoring. One could also argue that this is more subjective than other indices as the descriptions for each category

are relatively sparse and more reliant on experience and using the reference models during categorisation. Calibration may therefore be more important for this index than for the others in order to improve reliability.

The 5 Year Olds' index category descriptions are more comprehensive and less reliant on reference models, which would seem to be borne out by the results. Both the five and ten year age groups seemed to show similar reliability scores despite only five year reference models being available. Weighted Kappa scores were slightly higher overall compared to GOSLON, although the confidence intervals overlapped meaning there is less statistical evidence for a difference in reliability between the indices.

The modified Huddart/Bodenham index is promoted as not requiring any form of calibration prior to use due to its objectivity. The results of this study would seem to support this. The intra-examiner weighted Kappa scores were similar to the 5 Year Olds' index, but the inter-examiner scores were impressively high. This would certainly suggest a higher degree of objectivity, which would be especially beneficial to novice examiners.

The EUROCRAN index was slightly more difficult to analyse as this comprises two components. The dental component weighted Kappa scores performed similarly to the GOSLON Yardstick and 5 Year Olds' index, with the added benefit of having reference models available for both the five and ten year age groups. The EUROCRAN palatal component scores were the lowest of all the indices tested in this study, with the inter-examiner 5 year age group scores being particularly low at 0.51 and 0.553 (Table 2). This is similar to the findings of previous workers (17, 18). Palatal morphology is also an added complexity to the index. Although perhaps a relevant outcome when assessing surgical treatment, it may be better if it were incorporated into a final overall score looking at dental arch relationships, rather than being given its own separate score.

Overjet measurement reliability could only be measured meaningfully in the ten year age group, where it could be converted to a five point scale (no previous work exists to convert overjet at the other age ranges and this would require a separate study to produce a satisfactory conversion method). The weighted Kappa scores at this age group were similar to the 5 Year Olds' index and so would appear to be reliable. If it could be proved to be a valid outcome measure then further work to produce a method of conversion to five categories at the five year age group may be beneficial. However, validity of the conversion at the ten year age group is already questionable, as Morris et al., (12) used linear regression to produce an ordinal five point scale based on the overjet measurement, whereas it should strictly be used only on continuous normally distributed data.

Validity

True validity is not practical or ethical to assess as it would necessitate withholding treatment from patients from the moment of primary surgery to when they are fully grown in adulthood. The reality is the initial primary surgical outcome becomes distorted with subsequent surgical and orthodontic treatment as well as the patient's inherent growth pattern. Both methods of measuring validity in this study are compromises. Comparing indices to a gold standard is a well-recognised method of measuring validity. One method considered was comparing the index scores to independent expert consensus opinion. However, it was felt that certain indices (such as GOSLON) are so well established that they would influence an expert's opinion when trying to independently categorise a set of models. It was therefore decided that nominating a gold standard index at both the five and ten year age groups against which to compare the other indices was the best method. For the five year age group, the 5 Year Olds' index was selected as it was originally designed for this age group and is the most widely

used at this age (19-21). The GOSLON Yardstick was selected for the ten year age group for similar reasons (22-24). This does mean that no information is provided about the validity of these gold standard indices at these age groups (although this is partially addressed by also measuring predictive validity of all indices).

In the five year age group, the GOSLON Yardstick had an agreement weighted Kappa score of less than 0.6 with the gold standard index. Although this falls into the moderate agreement category, there is clearly a difference between the outcomes of the two indices. Similarly, at the ten year age group the 5 Year Olds' index had an agreement value of less than 0.7 to the gold standard GOSLON Yardstick. So, if there is a difference between the two indices at the two ages, which one should be used? This comes down to which is the most valid at each age group which, as mentioned earlier, is extremely difficult to prove. However, in the absence of scientific proof, common sense dictates that it may be more sensible to use each index in the age group for which it was originally designed and where there are reference models at the correct age for i.e. 5 Year Olds' index at five years and GOSLON Yardstick at ten years.

The modified Huddart/Bodenham index had a high agreement with the 5 Year Olds' index at five years of age, but had a poor agreement with the gold standard index in the ten year age group. Considering the high agreement in the five year age group, this was a disappointingly low agreement at the ten year age group. It may be that the conversion into the five categories needs to be improved in this age group or it may be that the examiner's GOSLON ratings are not accurate due to their lack of calibration. This latter point is only likely to account for a small part of this lack of agreement between the two indices based on the respectable reliability scores of the GOSLON Yardstick.

The palatal component of EUROCRAN recorded low correlation between the gold standard scores at both age groups. Although the dental component performed better, overall validity for this index must be questioned.

Measuring predictive validity provides information on all indices included in the study. The results were slightly underwhelming for all these indices. The predictive validity at the five year age group was very similar for GOSLON, 5 Year Olds' and modified Huddart/Bodenham indices. Around 50% of all of these indices scores stayed the same at their final outcomes. Interestingly, the GOSLON Yardstick seemed to judge outcomes less harshly than the 5 Year Olds' index, which in turn was less harsh than the modified Huddart/Bodenham index. The fact that only half of categorised models stayed in the same category highlights the difficulty in accurately predicting final outcome and need for orthognathic surgery in the future at such a young age. This finding supports previous work (5).

In the ten year age group, predictive validity for the above indices improved to above 60%. This improvement is perhaps unsurprising as more of the patient's growth pattern has been expressed by this time and so they will have less growth and treatment to come, which could alter the final outcome. Even with these improvements, one third of patients had a sufficiently large enough change in their growth pattern (or perhaps interim orthodontic treatment) to switch groups between ten and twenty years of age. Again, this finding supports previous work that future growth cannot be predicted based on 5 Year Olds' or GOSLON outcome (25). Overjet measurement at the ten year age group seemed to predict final outcome with less accuracy as only 44% of study models stayed in the same group at final outcome.

It is more difficult to draw direct comparison with the EUROCRAN index's predictive validity, but Table 3 shows that the dental component seems to correlate fairly well between the five and ten year age groups and the final outcome. The palatal component of the index seems to have very little correlation with the final outcome, although one could argue that the method of measuring final outcome in this study failed to account for palatal morphology and so it is perhaps unfair to draw too many conclusions from this.

Limitations

Conducting primary research in this field is extremely challenging and numerous compromises in the study design were necessary.

The five included indices do not all operate on the same scale, meaning that comparison between indices was complicated. Advice from the statistical department was taken and conversion of the indices to a five point scale was deemed to be the best approach. This allowed Kappas to be calculated for reliability which is the most commonly used and most familiar statistical test in the field.

Ideally, the examiners would have been very experienced in the use of each index included in the comparison. Using examiners with no experience in the indices yielded its own benefits of informing on ease of use without prior knowledge. Repeating the study using the best performing indices of GOSLON, 5 Year Olds' Index and modified Huddart/Boddenham with examiners similarly experienced in each index may be worthwhile in the future.

Collecting the sample proved very difficult. Finding cases meeting the inclusion criteria with study models at the correct age was the limiting factor. Using this study to inform on a power calculation for a future study may be beneficial in calculating the optimal sample size which

could be collected by searching a larger number of units. A sample size of around 30 for this purpose has been shown to be appropriate based on previous statistical work (26).

The age which the final models were collected was the biggest variation in the collected sample. The average age was 18 years and 2 months but some were as young as 15 years. It is likely that the 15 year olds have further growth to come after this age but unlikely that this would drastically alter the final outcome recorded at age 15 years.

Conclusions

- The GOSLON Yardstick proved to be simple to use as an outcome measure of primary UCLP surgery. Its use in the mixed dentition was more reliable and valid as opposed to the primary dentition would seem most appropriate considering the heavy reliance on the mixed dentition reference models during scoring.
- The 5 Year Olds' index was seemingly less reliant on reference models because of the thorough category descriptions. It was slightly more reliable than the GOSLON Yardstick, but was more time consuming to use. Its use in the primary dentition is recommended due to improved validity and allowing earlier audit of primary cleft surgery outcomes.
- The modified Huddart/Bodenham index proved to be the most reliable and objective primary surgery outcome measure, with claims of no calibration being necessary supported. It proved to be valid when used in the primary dentition. However, it was the most time consuming index to use and some questions remain over its validity in the mixed dentition.

- Neither the EUROCRAN index nor overjet measurement can be recommended as outcome measures of primary UCLP surgery because of unproven validity. EUROCRAN also had very low reliability scores.

- Prediction of final outcome at age 20 years was not reliable using any primary UCLP surgery outcome measure at either the five or ten year age group.

In summary, the results of this study support the use of the 5 Year Olds' Index and modified Huddart/Boddenham Index at 5 years of age, and GOSLON Yardstick at 10 years of age. There was no clear evidence to support one index at any one age group above all others.

Acknowledgements

This publication presents independent research commissioned by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research scheme (RP-PG-0707-10034). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

1. Ross, R.B. (1970) The clinical implications of facial growth in cleft lip and palate. *Cleft Palate J*, 7, 37-47.

2. Mars, M., Asher-McDade, C., Brattström, V., Dahl, E., McWilliam, J., Mølsted, K., et al. (1992) A six-center international study of treatment outcome in patients with clefts of the lip and palate: Part 3. Dental arch relationships. *Cleft Palate Craniofac J*, 29, 405-8.

3. Mars, M., Plint, D.A., Houston, W.J., Bergland, O., Semb, G. (1987) The Goslon Yardstick: a new system of assessing dental arch relationships in children with unilateral clefts of the lip and palate. *Cleft Palate J*, 24, 314-22.

4. Atack, N., Hathorn, I., Mars, M., Sandy, J. (1997) Study models of 5 year old children as predictors of surgical outcome in unilateral cleft lip and palate. *Eur J Orthod*, 19, 165-70.

5. Atack, N.E., Hathorn, I.S., Semb, G., Dowell, T., Sandy, J.R. (1997) A new index for assessing surgical outcome in unilateral cleft lip and palate subjects aged five: reproducibility and validity. *Cleft Palate Craniofac J*, 34, 242-6.

6. Mossey, P.A., Clark, J.D., Gray, D. (2003) Preliminary investigation of a modified Huddart/Bodenham scoring system for assessment of maxillary arch constriction in unilateral cleft lip and palate subjects. *Eur J Orthod*, 25, 251-7.

7. Gray, D., Mossey, P.A. (2005) Evaluation of a modified Huddart/Bodenham scoring system for assessment of maxillary arch constriction in unilateral cleft lip and palate subjects. *Eur J Orthod*, 27, 507-11.

8. Dobbyn, L.M., Weir, J.T., Macfarlane, T.V., Mossey, P.A. (2012) Calibration of the modified Huddart and Bodenham scoring system against the GOSLON/5-year-olds' index for unilateral cleft lip and palate. *Eur J Orthod*, 34, 762-767.

9. Fudalej, P., Katsaros, C., Bongaarts, C., Dudkiewicz, Z., Kuijpers-Jagtman, A.M. (2011) Dental arch relationship in children with complete unilateral cleft lip and palate following one-stage and three-stage surgical protocols. *Clin Oral Investig*, 15, 503-10.

10. Jones, T., Al-Ghatam, R., Atack, N., Deacon, S., Power, R., Albery, L., et al. (2014) A review of outcome measures used in cleft care. *J Orthod*, 41, 128-40.

11. Altalibi, M., Saltaji, H., Edwards, R., Major, P.W., Flores-Mir, C. (2013) Indices to assess malocclusions in patients with cleft lip and palate. *Eur J Orthod*, 35, 772-782.

12. Morris, T., Roberts, C., Shaw, W.C. (1994) Incisal overjet as an outcome measure in unilateral cleft lip and palate management. *Cleft Palate Craniofac J*, 31, 142-5.

13. Johnson, N., Williams, A.C., Singer, S., Southall, P., Atack, N., Sandy, J.R. (2000) Dentoalveolar relations in children born with a unilateral cleft lip and palate (UCLP) in Western Australia. *Cleft Palate Craniofac J*, 37, 12-6.

14. Mølsted, K., Brattström, V., Prahl-Andersen, B., Shaw, W.C., Semb, G. (2005) The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 3: dental arch relationships. *Cleft Palate Craniofac J*, 42, 78-82.

15. Lilja, J., Mars, M., Elander, A., Enocson, L., Hagberg, C., Worrell, E., et al. (2006) Analysis of dental arch relationships in Swedish unilateral cleft lip and palate subjects: 20-year longitudinal consecutive series treated with delayed hard palate closure. *Cleft Palate Craniofac J*, 43, 606-11.

16. Hathaway, R., Daskalogiannakis, J., Mercado, A., Russell, K., Long, R.E., Cohen, M., et al. (2011) The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 2. Dental arch relationships. *Cleft Palate Craniofac J*, 48, 244-51.

17. Fudalej, P., Katsaros, C., Dudkiewicz, Z., Offert, B., Piwowar, W., Kuijpers, M., et al. (2012) Dental arch relationships following palatoplasty for cleft lip and palate repair. *J Dent Res*, 91, 47-51.

18. Patel, D. (2011) Evaluation of the use of the Modified Huddart Boddenham & Eurocran Yardstick for the assessment of surgical outcome for unilateral cleft lip and palate, University of Dundee.

19. Williams, A.C., Johnson, N.C., Singer, S., Southall, P., Mildinhall, S., Semb, G., et al. (2001) Outcomes of cleft care in Western Australia: a pilot study. *Aust Dent J*, 46, 32-6.

20. Flinn, W., Long, R.E., Garattini, G., Semb, G. (2006) A multicenter outcomes assessment of five-year-old patients with unilateral cleft lip and palate. *Cleft Palate Craniofac J*, 43, 253-8.

21. Clark, S.A., Atack, N.E., Ewings, P., Hathorn, I.S., Mercer, N.S. (2007) Early surgical outcomes in 5-year-old patients with repaired unilateral cleft lip and palate. *Cleft Palate Craniofac J*, 44, 235-8.

22. Johnston, C.D., Leonard, A.G., Burden, D.J., McSherry, P.F. (2004) A comparison of craniofacial form in Northern Irish children with unilateral cleft lip and palate treated with different primary surgical techniques. *Cleft Palate Craniofac J*, 41, 42-6.

23. Liao, Y.F., Lin, I.F. (2009) Dental arch relationships after two-flap palatoplasty in Taiwanese patients with unilateral cleft lip and palate. *Int J Oral Maxillofac Surg*, 38, 1133-6.

24. Jack, H.C., Antoun, J.S., Fowlert, P.V. (2011) Evaluation of primary surgical outcomes in New Zealand patients with unilateral clefts of the lip and palate. *Aust Orthod J*, 27, 23-7.

25. Suzuki, A., Sasaguri, M., Hiura, K., Yasunaga, A., Mitsuyasu, T., Kubota, Y., et al. (2014) Can Occlusal Evaluation of Children With Unilateral Cleft Lip and Palate Help Determine Future Maxillofacial Morphology? *Cleft Palate Craniofac J*, 51, 696-706.

26. Browne, R.H. (1995) On the use of a pilot sample for sample size determination. *Stat Med*, 14. 1933-40.

Figure Legend

Figure 1. Profile view of the 5 Year Olds' Index reference models.

Figure 2. Typical layout of study models during scoring sessions.

Figure 3. Mean scoring time (95% reference range illustrated as arrows) per index for 5 and 10 year age groups in index comparison study.

Figure 4. Level of agreement of GOSLON and modified Huddart/Bodenham to the gold standard 5 Year Olds' index for scoring the five year age group in the index comparison study.

Figure 5. Level of agreement of 5 Year Olds' index, modified Huddart/Bodenham and overjet measurement to the gold standard GOSLON Yardstick for scoring the ten year age group in the index comparison study.

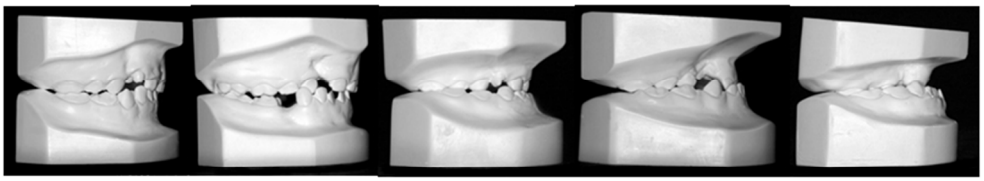
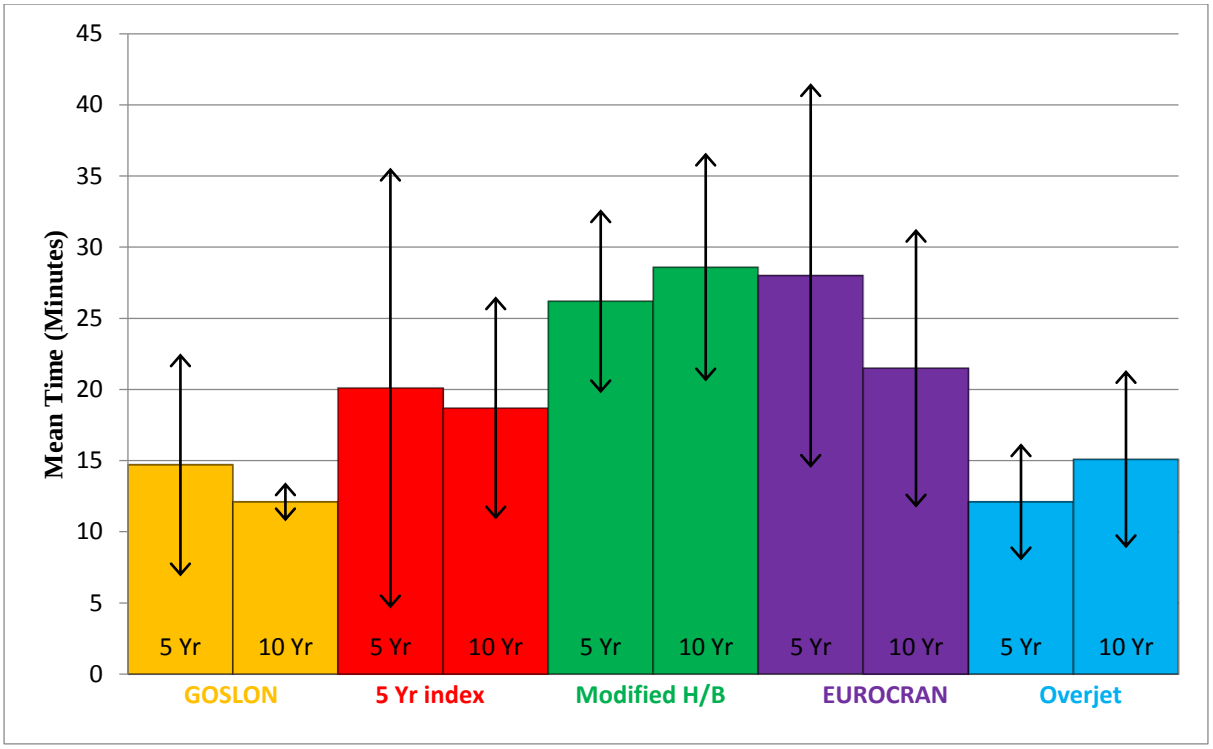
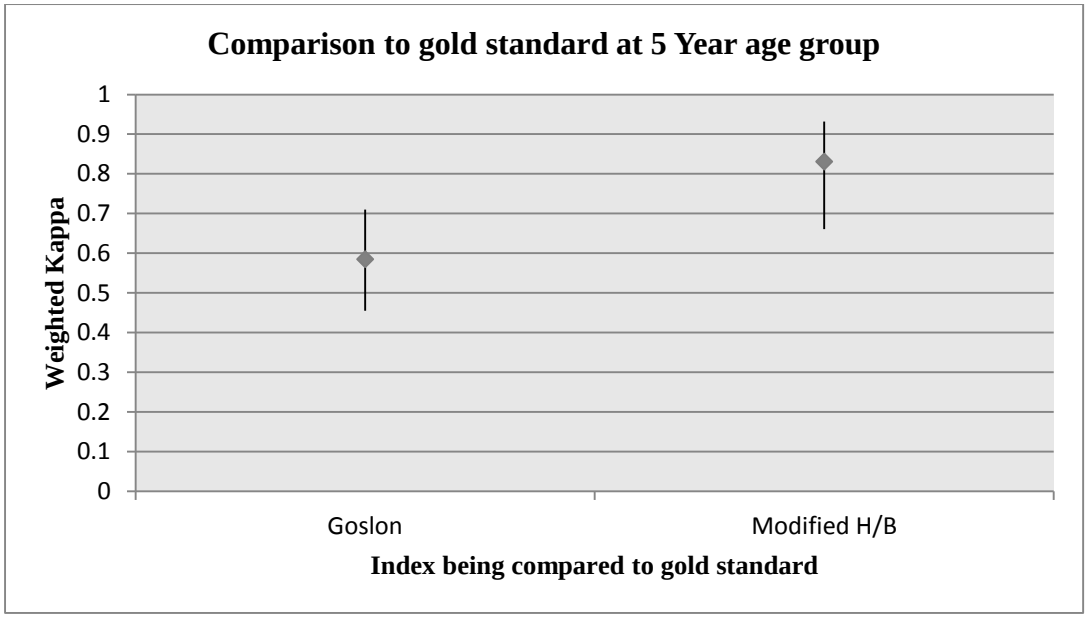
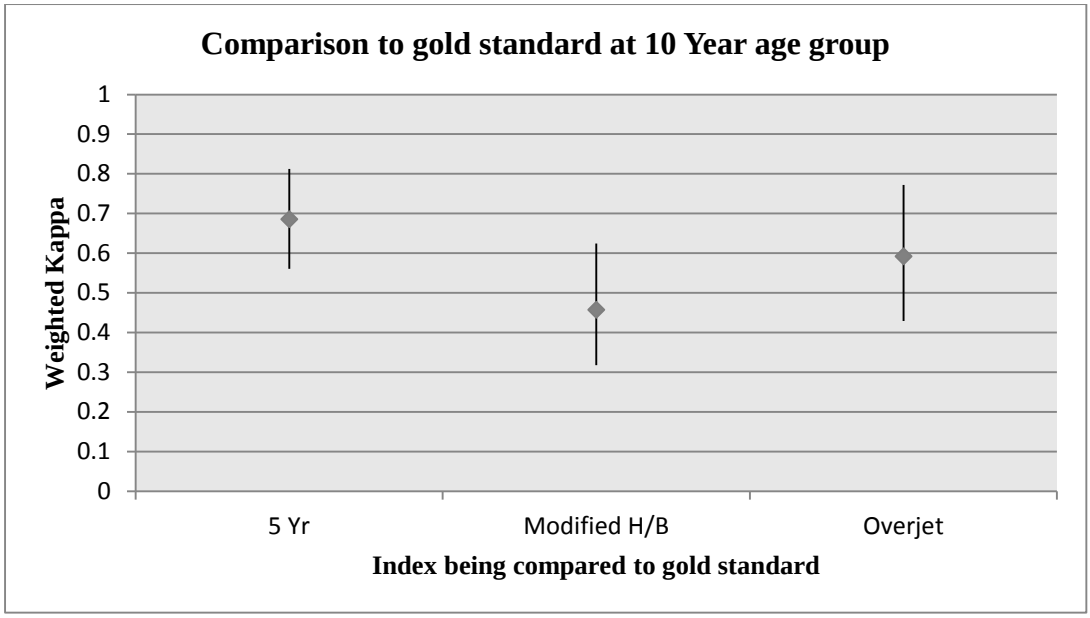


Figure 1. Profile view of the 5 Year Olds' Index reference models.
216x56mm (96 x 96 DPI)

Figure 2. Typical layout of study models during scoring sessions.
130x97mm (220 x 220 DPI)


Mean Time (Minutes)
45
40
35
30
25
20
15
10
5
0
5 Yr
10 Yr
5 Yr
10 Yr
5 Yr
10 Yr
5 Yr
10 Yr
5 Yr
10 Yr
GOSLON
5 Yr index
Modified H/B
EUROCRAN
Overjet


Comparison to gold standard at 5 Year age group
Weighted Kappa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Goslon
Modified H/B
Index being compared to gold standard


Comparison to gold standard at 10 Year age group
Weighted Kappa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5 Yr
Modified H/B
Overjet
Index being compared to gold standard

Questionnaire given to examiners to gather opinion on the different indices for the index comparison

Questionnaire on UCLP indices used to score primary surgical outcome on study models

Please complete for each index following completion of scoring session two

Assessor name:

Please circle the index which your comments relate to:

Eurocran index 5 Year Olds' index Goslon Yardstick Modified Huddart Bodenham index

OJ Measurement

How easy did you find scoring the study models using this index (please circle)?

Very difficult

Very easy

1 2 3 4 5 6 7 8 9 10

What do you feel the positive aspects of this index are?

What do you feel the negative aspects of this index are?

Records	Age range	Mean age	1 Std deviation
5 year	4yrs 9m to 6yrs 8m	5 yrs, 3 months	7.9 months
10 year	8yrs 0m to 12yrs 5m	9 yrs, 11 months	1 yr, 5.2 months
Post-treatment	15yrs 0m to 26 yrs 5m	18 yrs, 2 months	2 yrs, 3.8 months

Table 1. Mean age and standard deviation of patients included in index comparison sample when each set of study models were taken.

Age group	Examiner	Weighted Kappa (95% confidence interval) per index					
		GOSLON	5 Year Olds'	Modified H/B	Eurocran dental	Eurocran palatal	Overjet
5 year	K.J.D	0.52 (0.39 to 0.67)	0.87 (0.76 to 0.96)	0.77 (0.62 to 0.91)	0.81 (0.68 to 0.91)	0.76 (0.58 to 0.90)	-
	R.R	0.75 (0.60 to 0.88)	0.71 (0.57 to 0.85)	0.71 (0.56 to 0.85)	0.74 (0.55 to 0.87)	0.73 (0.54 to 0.91)	-
10 year	K.J.D	0.86 (0.71 to 0.97)	0.90 (0.80 to 0.96)	0.91 (0.81 to 0.97)	0.79 (0.64 to 0.91)	0.86 (0.68 to 0.97)	0.90 (0.74 to 1.00)
	R.R	0.70 (0.52 to 0.86)	0.71 (0.58 to 0.83)	0.87 (0.73 to 0.96)	0.54 (0.39 to 0.71)	0.68 (0.48 to 0.86)	0.74 (0.56 to 0.90)
Final	K.J.D	0.95 (0.86 to 1.00)	-	-	-	-	-
	R.R	0.75 (0.54 to 0.89)	-	-	-	-	-

Table 2. Index comparison intra-examiner kappa scores.

Age group	Scoring session	Weighted Kappa (95% confidence interval) per index					
		GOSLON	5 Year Olds'	Modified H/B	Eurocran dental	Eurocran palatal	Overjet
5 year	1	0.41 (0.25 to 0.55)	0.76 (0.61 to 0.88)	0.81 (0.66 to 0.91)	0.75 (0.58 to 0.88)	0.55 (0.35 to 0.74)	-
	2	0.65 (0.49 to 0.81)	0.83 (0.70 to 0.90)	0.79 (0.63 to 0.91)	0.76 (0.60 to 0.90)	0.51 (0.31 to 0.71)	-
10 year	1	0.68 (0.50 to 0.85)	0.70 (0.55 to 0.83)	0.91 (0.81 to 0.96)	0.70 (0.54 to 0.85)	0.70 (0.47 to 0.87)	0.86 (0.69 to 0.97)
	2	0.70 (0.55 to 0.88)	0.75 (0.61 to 0.86)	0.83 (0.70 to 0.93)	0.67 (0.51 to 0.82)	0.70 (0.49 to 0.88)	0.72 (0.52 to 0.88)
Final	1	0.68 (0.50 to 0.82)	-	-	-	-	-
	2	0.65 (0.39 to 0.80)	-	-	-	-	-

Table 3. Index comparison inter-examiner kappa scores.

Age group	Index	Comparison to 20 year age group	Percentage
Five year age group	GOSLON Yardstick	Stayed the same	50.00
		Improved	23.53
		Worsened	26.47
	5 Year Olds' index	Stayed the same	52.94
		Improved	32.35
		Worsened	14.71
	Modified Huddart/Bodenham	Stayed the same	50.00
		Improved	32.35
		Worsened	17.65
Ten year age group	GOSLON Yardstick	Stayed the same	64.71
		Improved	17.65
		Worsened	17.65
	5 Year Olds' index	Stayed the same	64.71
		Improved	23.53
		Worsened	11.76

	Modified Huddart/Bodenham	Stayed the same	61.76
		Improved	32.35
		Worsened	5.88
	Overjet measurement	Stayed the same	44.12
		Improved	32.35
		Worsened	23.53

Table 4. Percentage of cases which scored the same, better or worse at the twenty year age group compared to the scores given at the 5 and 10 year age group.

Index and age group	Spearman's correlation coefficient (p value)
EUROCRAN dental, age group = 5	0.45 (0.008)
EUROCRAN palatal, age group = 5	0.21 (0.244)
Overjet, age group = 5	-0.39 (0.023)
EUROCRAN dental, age group = 10	0.57 (0.001)
EUROCRAN palatal, age group= 10	0.20 (0.256)

Table 5. Table illustrating the Spearman's correlation coefficients comparing the five indices at 5 and 10 years with the final outcome at 20 years. P value in brackets testing the null hypothesis that there is no correlation between EUROCRAN/overjet measurement and final outcome.

Age group	Index	Ease of use (1-10, very difficult-very easy)
5 yrs	GOSLON Yardstick	6.3
	5 Year Olds' Index	7
	Modified Huddart/Bodenham	6.5
	EUROCRAN Index	3.25
	Overjet measurment	8
10 yrs	GOSLON Yardstick	7.5
	5 Year Olds' Index	6
	Modified Huddart/Bodenham	6.5
	EUROCRAN Index	3.5
	Overjet measurment	8.5

Table 6. Average ease of use subjective scores assigned by examiners after scoring with each index at each age group.